Apache Pig Tutorial

TSA Workshop

Pranamesh Chakraborty

Resources: CPRE-419 course in ISU (Large Scale Data Analysis)

Apache Pig

- Framework for large scale data processing, at a higher level of abstraction than MapReduce.

- Writes programs faster than MapReduce for processing large datasets

Apache Pig

Faster Development

Write a Program for Word Count using only 5 lines of code.

Example: count number of occurrences of each word in a corpus stored in ‘words.txt’

```
shak = LOAD 'words.txt' AS (line);

all_words = FOREACH shak GENERATE
FLATTEN(TOKENIZE(line)) AS word;

word_grps = GROUP all_words BY word;

counts = FOREACH word_grps GENERATE group,
COUNT(word);

STORE counts INTO ’wc_output';
```

HDFS Login

1. Follow instructions given in Github

Common hdfs commands

Starts with *hdfs dfs -….*
or *hadoop fs -….*

- See the contents of a folder:
  - *hdfs dfs –ls <location>*
  - Examples:

hdfs dfs –ls

hdfs dfs –ls inrix

Common hdfs commands

- Make a new directory in hdfs:
  - *hdfs dfs –mkdir <hdfs directory location>*

```
hdfs dfs –mkdir pranamesh
```

- See the tail of a file in hdfs:
  - *hdfs dfs –tail <hdfs file location>*

```
hdfs dfs –tail inrix/2-2-2017.csv
```

- See the top of a file in hdfs:
  - *hdfs dfs –cat <hdfs file name>|head -10*

```
hdfs dfs –cat inrix/2-2-2017.csv|head -10
```

Common hdfs commands

- Copy to Local machine from HDFS
  - *hdfs dfs –copyToLocal <local machine location> <location in HDFS>*
  - Then copy the required file from the local machine to your machine via WinScp

```
hdfs dfs –copyToLocal inrix/sample.csv pranamesh
```

Common hdfs commands

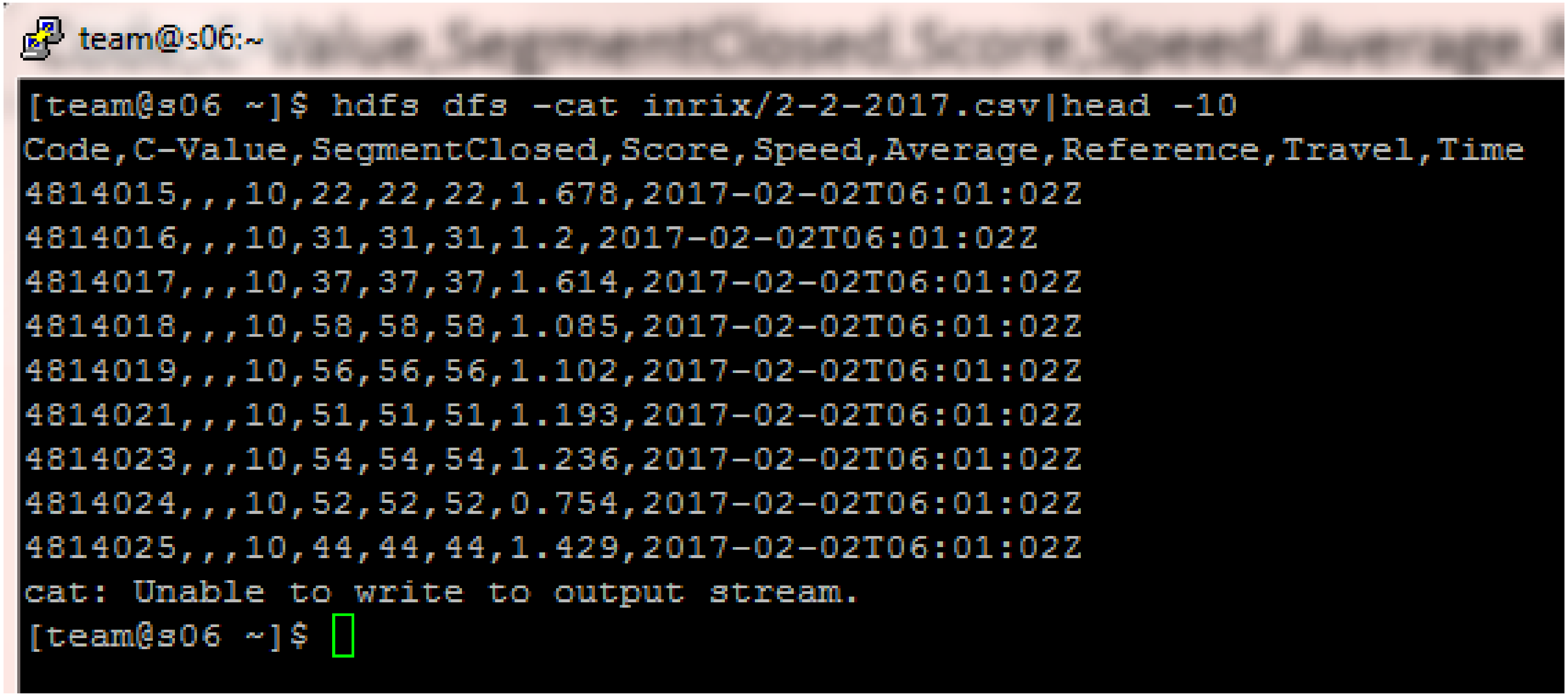
- Copy from Local machine to HDFS
  - First copy the required file to the local machine via WinScp
  - hdfs dfs –copyFromLocal <local machine location> <location in HDFS>

```
hdfs dfs –copyFromLocal pranamesh/sample.csv pranamesh
```

Pig Script

A sample script on INRIX XD Data

Inrix XD data Schema:
Code,C-Value,SegmentClosed,Score,Speed,Average,Reference,Traveltime,Time



```
[team@s06 ~]$ hdfs dfs -cat inrix/2-2-2017.csv|head -10
Code,C-Value,SegmentClosed,Score,Speed,Average,Reference,Travel,Time
4814015,,,10,22,22,22,1.678,2017-02-02T06:01:02Z
4814016,,,10,31,31,31,1.2,2017-02-02T06:01:02Z
4814017,,,10,37,37,37,1.614,2017-02-02T06:01:02Z
4814018,,,10,58,58,58,1.085,2017-02-02T06:01:02Z
4814019,,,10,56,56,56,1.102,2017-02-02T06:01:02Z
4814021,,,10,51,51,51,1.193,2017-02-02T06:01:02Z
4814023,,,10,54,54,54,1.236,2017-02-02T06:01:02Z
4814024,,,10,52,52,52,0.754,2017-02-02T06:01:02Z
4814025,,,10,44,44,44,1.429,2017-02-02T06:01:02Z
cat: Unable to write to output stream.
[team@s06 ~]$
```

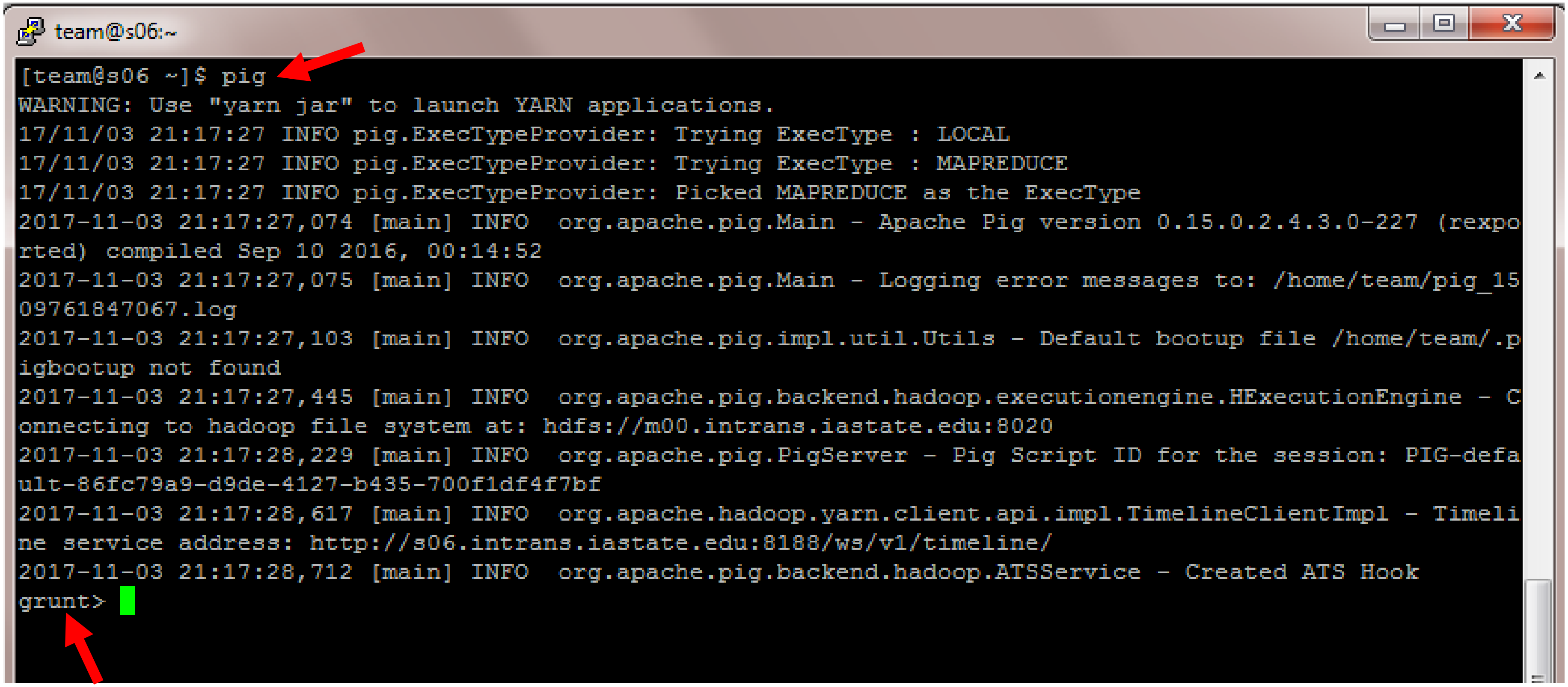
Pig Script

Inrix XD data Schema:
Code,C-Value,SegmentClosed,Score,Speed,Average,Reference,Traveltime,Time

- **Problem:** Count the number of occurrences of confidence score = 30 for any 10 segments with the sample Inrix XD data, and output them to a file

Pig Grunt Shell

Type “pig” in your putty terminal. The grunt shell is open.

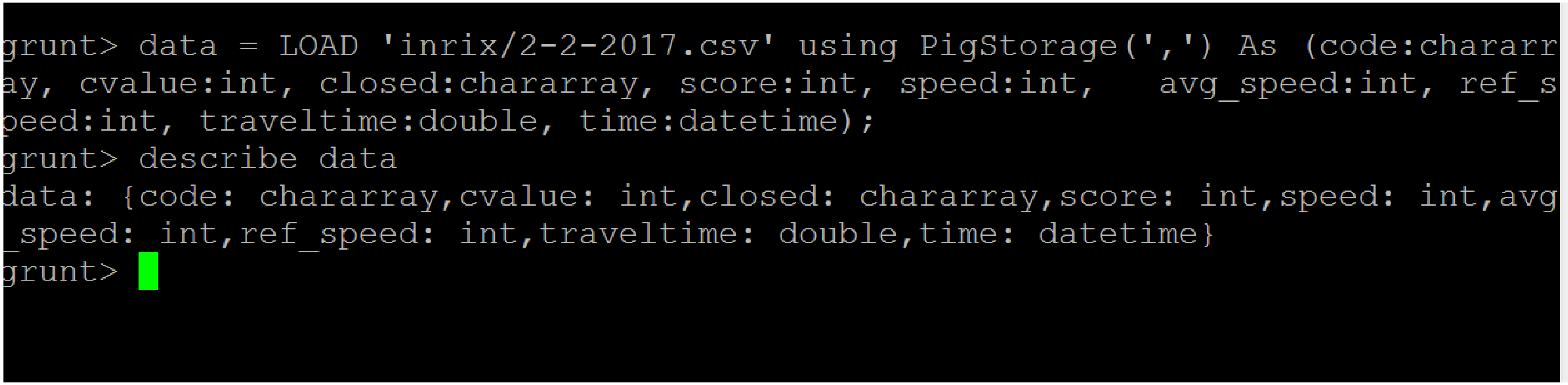


```
[team@s06 ~]$ pig
WARNING: Use "yarn jar" to launch YARN applications.
17/11/03 21:17:27 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
17/11/03 21:17:27 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
17/11/03 21:17:27 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-11-03 21:17:27,074 [main] INFO  org.apache.pig.Main - Apache Pig version 0.15.0.2.4.3.0-227 (rexpo
rted) compiled Sep 10 2016, 00:14:52
2017-11-03 21:17:27,075 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/team/pig_15
09761847067.log
2017-11-03 21:17:27,103 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/team/.p
igbootup not found
2017-11-03 21:17:27,445 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - C
onnecting to hadoop file system at: hdfs://m00.intrans.iastate.edu:8020
2017-11-03 21:17:28,229 [main] INFO  org.apache.pig.PigServer - Pig Script ID for the session: PIG-defa
ult-86fc79a9-d9de-4127-b435-700f1df4f7bf
2017-11-03 21:17:28,617 [main] INFO  org.apache.hadoop.yarn.client.api.impl.TimelineClientImpl - Timeli
ne service address: http://s06.intrans.iastate.edu:8188/ws/v1/timeline/
2017-11-03 21:17:28,712 [main] INFO  org.apache.pig.backend.hadoop.ATSService - Created ATS Hook
grunt> 
```

Pig Grunt Shell

Type each of the commands given in Github pig script in the grunt shell

Script location: https://github.com/pranamesh/Python-workshop-TSA/blob/master/Apache-Pig/test_script.pig



```
grunt> data = LOAD 'inrix/2-2-2017.csv' using PigStorage(',') As (code:chararr
ay, cvalue:int, closed:chararray, score:int, speed:int,   avg_speed:int, ref_s
peed:int, traveltime:double, time:datetime);
grunt> describe data
data: {code: chararray,cvalue: int,closed: chararray,score: int,speed: int,avg
_speed: int,ref_speed: int,traveltime: double,time: datetime}
grunt> 
```

Pig Script

Run the script:

```
pig <script location in Local machine>
```

```
pig Pranamesh/test-script.pig
```

Apache Pig

Resources:

Reference Manual:

https://pig.apache.org/docs/r0.7.0/piglatin_ref2.html

Built-in functions

https://pig.apache.org/docs/r0.11.1/func.html